



2006-2007 Reporting Guide for Maine Educational Assessment (MEA) Scores

Reporting Guide for Maine Educational Assessment (MEA) Scores

Scaled scores are the most suitable statistic to use when comparing MEA results among students, schools, and districts. When interpreting the meaning of these comparisons, however, it is important that decision-makers—teachers, administrators, and policy-makers—fully recognize that any single test is a limited measure of student performance. Since some apparent differences in scaled scores may not be statistically or educationally significant, some guidelines for comparing results are explained on the following pages.

COMPARISONS OF SCHOOL- AND DISTRICT-LEVEL SCORES

The statistical significance of these comparisons is based on variability of the scores and on the number of students tested. The table on the following page can be used to assist you in the following ways:

- comparing sub-populations of students within a school or district,
- comparing the scores of two or more schools or districts,
- comparing the scores of a school to the district and/or state, and
- comparing the scores of a district to the state.

These tables provide figures that can be used to make approximate comparisons between scores. Similar to the score band provided in the student report, the figures in the tables are estimates of one standard error of measurement around the score or difference between scores. Caution should be used when making any of the comparisons listed above because even if scores are different they may not be statistically significantly different. It is very unlikely that any two groups, even if they have similar levels of achievement, will have exactly the same score. To avoid misinterpretation or over-interpretation of small differences between scores, statistical tests can be conducted to determine the likelihood that the observed difference in scores occurred by chance and that the two groups might actually have the same score. A difference is statistically significant if one can be 95% sure that the difference is not due to chance.

SCALED SCORES

MEA scaled scores are represented by a 3-digit number, with the first digit representing the grade level tested; the remaining digits range from 00–80. Although this same scale is used for all content areas, one cannot accurately compare a school's or district's scaled scores across two content areas since the scaled scores in each content area were determined by separate standard-setting processes. The table on the following page shows the smallest differences in scaled scores that represent a statistically significant difference in performance based on the number of students tested in the school and/or district. When comparing the scores of two groups of different sizes, one should use a difference that is approximately the average of the minimally statistically significant difference of each group. For example, when comparing the average Grade 3 reading scaled score of a school with 25 students to the average Grade 3 reading scaled score of a school with 100 students, one should use two points as the minimally statistically significant difference. Two points is the average of the values for a school of 25 students (3 points) and a school of 100 students (1 point). If the difference in scaled scores between the two groups is at least two points, then the difference is statistically significant. If the difference in scaled scores between the two groups is fewer than two points, the difference is not statistically significant.

Number of Scaled Score Points Denoting Minimally Statistically Significant Differences for Average School/District Results						
Grade	Subject	Number of Students Tested in School/District				
		10	25	50	100	200
3	Reading	4	3	2	1	1
	Mathematics	7	4	3	2	1
4	Reading	4	3	2	1	1
	Mathematics	6	4	3	2	1
	Science	6	4	3	2	1
5	Reading	5	3	2	2	1
	Writing	5	3	2	2	1
	Mathematics	7	4	3	2	1
6	Reading	6	3	2	2	1
	Mathematics	8	5	3	2	2
7	Reading	7	4	3	2	1
	Mathematics	8	5	4	3	2
8	Reading	7	4	3	2	2
	Writing	6	4	3	2	1
	Mathematics	8	5	4	3	2
	Science	6	4	3	2	1

Standard error of the mean difference with assumption $n_1=n_2$ and $s_1=s_2$.

ACHIEVEMENT LEVELS

Comparisons of group performance can also be made by comparing the percentages of students scoring at or above a particular achievement level. But again, small differences in percentages should not be over-interpreted. Because, unlike scaled scores, achievement level results are reported as percentages, a slightly different procedure is used to make comparisons between the performance of two groups or between a group and a fixed point. To compare percentages, an interval estimation approach similar to a margin of error or the score band reported on the MEA student report can be used.

With percentages, the statistical significance of differences is impacted by both the size of the group and the percentage of students in the category of interest (for example, Meets or above on the Grade 4 Mathematics test). The table on the following page shows the size of the confidence interval that should be drawn around a score for selected percentages and school sizes. For example, if 60% of the students in a school of 50 students are at Meets or above, a confidence interval of ± 14 percentage points, from 46% to 74%, would be drawn around the score of 60%. If the school's performance were being compared to a fixed percentage of 65% of students at Meets or above, the conclusion would be that the school score was not significantly different because the 46%-74% confidence interval includes 65%.

Percentage Difference in Student Achievement Level Classifications Denoting Minimally Statistically Significant Differences for School/District Results					
Percentages of Students in Category	Number of Students Tested in School District				
	10	25	50	100	200
10	19	12	8	6	4
20	25	16	11	8	6
30	28	18	13	9	6
40	30	19	14	10	7
50	31	20	14	10	7
60	30	19	14	10	7
70	28	18	13	9	6
80	25	16	11	8	6
90	19	12	8	6	4

*.05 level of statistical significance

The previous example compared the performance of a relatively small school to a fixed point (for example, a very large group such as the state). When two relatively small groups are compared, a confidence interval should be drawn around each score using the appropriate values from the table based on the size and performance of each group. If the two confidence intervals do not overlap, then the conclusion is that the difference between the two groups is statistically significantly different. If the two confidence intervals do overlap, then the difference in performance between the two groups is too small to be considered statistically significant. The distance between the two confidence intervals or their degree of overlap also provides a visual indication of the probability that the two scores are significantly different.

COMPARISONS OF MEA SCORES ACROSS YEARS

The comparison of scores across years requires consideration and caution in addition to those described in the previous section. In general, the evaluation of any score differences should always be interpreted within the larger context of what occurred to impact the performance of the school, district, or other group between the two test administrations being compared. Current MEA scores should not be compared to scores prior to 2006 because of the standard setting revision in 2006.

SCHOOL- AND DISTRICT-LEVEL SCALED SCORES AND ACHIEVEMENT LEVELS

The comparison of school- and district-level scaled scores and achievement levels across years is essentially the same as the comparison of similar scores within years. The procedures and cautions described in the previous section can be applied to scores from different years. As stated above, however, the interpretation of differences between scores should include consideration of any intervening factors between test administrations. Also note that when interpreting changes in performance across years, it can be beneficial to consider scaled scores and achievement levels jointly. Interpreting scaled scores or achievement levels alone may lead to misinterpretation or over-interpretation of results. Consider these examples:

- It is not unusual for large numbers of students to earn the same scaled score – particularly in the middle of the distribution near the Partially Meets/Meets cut

score. Consequently, school results may show a very small change in mean scaled score near the Meets cut score, but show a shift of 4-6 percentage points in the percentage of students performing at the Meets level or above.

- Conversely, a significant change in mean scaled score in the middle of an achievement level may not be reflected in improvement in the achievement level results.

STUDENT-LEVEL SCALED SCORES AND ACHIEVEMENT LEVELS

With MEA testing at grades 3 through 8, most students will have multiple years of MEA test scores. A logical question to ask is how the student's performance this year compares to performance in previous years. The most direct comparison can be made between a student's achievement level from one year to the next within a content area. The MEA tests are designed specifically to measure the grade level expectations for each grade. Students meeting or exceeding those expectations at their grade level should score at the Meets or Exceeds level each year. Of course, scores from a single test such as the MEA tests should always be interpreted with caution.

The question of whether student performance Meets or Exceeds the Standard at a particular grade level is critical, but we may also wish to examine progress toward proficiency within an achievement level. Achievement levels and scaled scores can be used together to examine, at a slightly finer level, whether a student is making progress toward proficiency from one year to the next. Scaled scores provide information about student performance within each achievement level. MEA scores are reported on separate 80-point scales corresponding to each grade level (300-380, 400-480, ..., 800-880). Each individual grade level scale has been developed so that at every grade a score of 41 represents performance that Meets the Standard at that grade level.

Although the tests and scales are different at each grade level, in general, for students performing below the Meets level, progress toward proficiency can be shown by earning a score that is closer to the Meets score of 41. For students scoring at the Does not Meet level, progress can be shown by earning a scaled score the next year that is closer to or within the Partially Meets level. Similarly, students scoring above Meets can progress toward the Exceeds level.

Of course, small differences in scores of 2-4 points on the 80-point scale should not be over-interpreted. As indicated by the score band on the student report, an individual score should be interpreted as a probable range of scores within which student performance might fall. For example, if a student earns a score of 438 in the fourth grade and 542 in the fifth grade, it is likely that the score bands for both grades will cross the Meets scores of 441 and 541, respectively. Therefore, the scores of 438 and 542 should not be considered significantly different from each other in relation to the Meets standard for these two grade levels. It is important to remember, however, that maintaining a score on the 80 point scale from one grade level to the next demonstrates approximately one year of growth on the content standards for each grade in that content area.

CONTENT AREA SUBSCORES

Content area subscores cannot be directly compared from one year to the next even within a grade level. Unlike achievement levels and scaled scores, these scores are reported as raw scores and have not been linked across years and placed on the same scale. Differences in

subscores from one year to the next in the total number of points earned by a student or in the percent of total possible points earned by a school or district may simply reflect either a small difference in the number of possible points in the reporting category or a slight difference in the difficulty of items within a particular reporting category. The process of equating that accounts for these differences to produce scaled scores and achievement levels for the total content area is not applied to individual reporting categories. There is not a sufficient number of points within each reporting category to equate these subscores from one year to the next.

There are, however, comparisons that can be made with content area subscores to assist schools in the evaluation of their curricula and instructional programs. For each content area subscore, normative information is provided describing performance in comparison to the school, district, state, and at the student level, students scoring at the Meets threshold. Across years, this information can be used to determine whether progress has been made relative to one of the comparison groups. Even more than with scaled scores and achievement levels, it is important not to over-interpret small changes from one year to the next.

At the school- and district-levels, it is also possible to pool content area subscores across years to compute a cumulative total beginning in 2006. Consistent with the cumulative achievement level and scaled score information reported for the total content area, results based on a larger pool of students and/or test items can provide a more stable picture of school or district performance over longer periods of time. Of course, intervening factors such as program or curricular changes may impact local decisions on the appropriateness of pooling data across years.